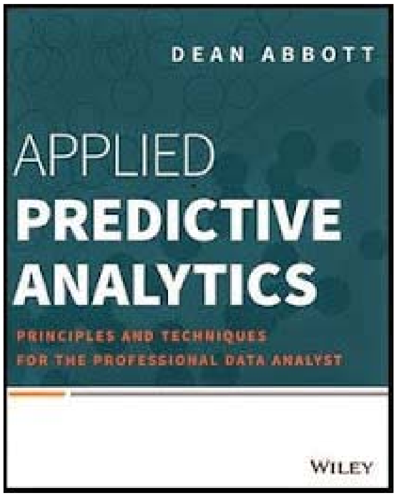
EXHIBIT 29

G+ More ▾ Next Blog» Create Blog Sign In

DATA MINING AND PREDICTIVE ANALYTICS

TIPS, TRICKS, AND COMMENTS IN DATA MINING AND PREDICTIVE ANALYTICS, INCLUDING DATA PREPROCESSING, VISUALIZATION, MODELING, AND MODEL DEPLOYMENT.
HOSTED BY DEAN ABBOTT, ABBOTT ANALYTICS, INC.

APPLIED PREDICTIVE ANALYTICS



CONTRIBUTORS

Dean Abbott
Will Dwinnell

OUR WEB SITES

Abbott Analytics, Inc.
Will's Data Mining in MATLAB

TWITTER UPDATES 2.2: FEEDWITTER

THURSDAY, MAY 01, 2014

Why Overfitting is More Dangerous than Just Poor Accuracy, Part I

Arguably, the most important safeguard in building predictive models is complexity regularization to avoid overfitting the data. When models are overfit, their accuracy is lower on new data that wasn’t seen during training, and therefore when these models are deployed, they will disappoint, sometimes even leading decision makers to believe that predictive modeling “doesn’t work”.

Overfit, however, is thankfully a well-known problem and every algorithm has ways to avoid it. CART® and C5 trees use pruning to remove branches that are prone to overfitting, CHAID trees require splits are statistically significant to add complexity to the trees. Neural networks use held-out data to stop training when accuracy on held-out data becomes worse. Stepwise regression uses information theoretic criteria like the Akaike Information Criterion (AIC), Minimum Description Length (MDL), or the Bayesian Information Criterion (BIC) to add terms only when the additional complexity is offset by enough reduction of error.

But overfitting has more problems than merely misclassification cases in holdout data or incurring large errors for regression problems. Without loss of generality, this discussion will only describe overfilling in classification problems, but the same principles apply in regression problems as well.

One way modelers reduce the likelihood of overfit is to apply the principle of Occam’s Razor, where if two models exhibit the same accuracy, we will prefer the simpler model because it is more likely to generalize well. By simpler, we must keep in mind that we prefer models that *behave* more simply rather than models that just appear to be simpler because they have fewer terms. John Elder (a regular contributor to the PA Times) has a fantastic discussion of

SUBSCRIBE TO THIS BLOG

Posts

Comments

SMART DATA COLLECTIVE

SmartData Collective Contributor

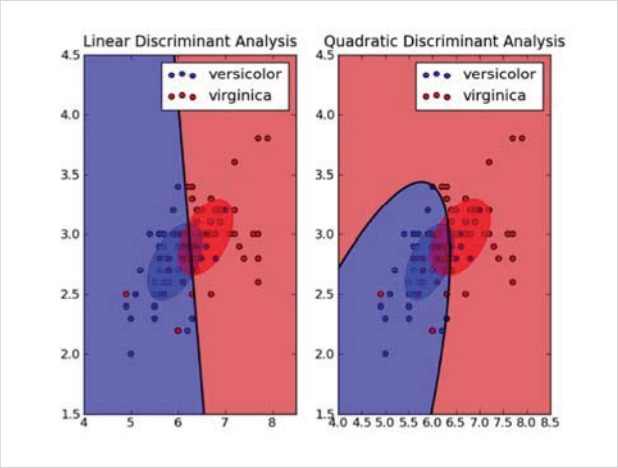
BLOG ARCHIVE

- ▶ 2017 (2)
- ▶ 2016 (1)
- ▶ 2015 (2)
- ▼ 2014 (5)
  - ▶ July (1)
  - ▼ May (2)
    - Why Overfitting is More Dangerous than Just Poor A...
    - Why Overfitting is More Dangerous than Just Poor A...
  - ▶ January (2)
- ▶ 2013 (13)
- ▶ 2012 (17)
- ▶ 2011 (17)
- ▶ 2010 (32)
- ▶ 2009 (23)
- ▶ 2008 (14)
- ▶ 2007 (42)
- ▶ 2006 (18)
- ▶ 2005 (6)
- ▶ 2004 (1)
- ▶ 2003 (1)

that topic in the book by Seni and Elder, Ensemble Methods in Data Mining.

Consider this example contrasting linear and nonlinear models. The figure below shows decision boundaries for two models separates two classes of the famous Iris Data (http://archive.ics.uci.edu/ml/datasets/Iris). On the left is the decision boundary from a linear model built using linear discriminant analysis (like LDA or the Fisher Discriminant) and on the right, a decision boundary built by a model using quadratic discriminant analysis (like the Bayes Rule). The image can be found at http://scikit-learn.org/0.5/auto_examples/plot_lda_vs_qda.html.

It appears that the accuracy of both models is the same (let's assume that it is), yet the behavior of the models is very different. If there is new data to be classified that appears in the upper left of the plot, the LDA model will call the data point versicolor whereas the QDA model will call it virginica. Which is correct? We don't know which would be correct from the training data, but we do know this: there is no justification in the data to increase the complexity of the model from linear to quadratic. We probably would prefer the linear model here.



Apply models to regions in the data without data is the entire reason for avoiding overfit. The issue with the figure above was with model behavior when doing *extrapolation*, where we want to make sure that the models behave in a reasonable way for values outside (larger than or smaller than) the data used in training. But models also need to behave well when they *interpolate*, meaning

DATA MINING BLOGS AND SITES

Data Mining in Matlab -- Will Dwinnell
Tim Mann's Data Mining Blog
Data Mining Research
Kevin Hillstrom's MineThatData
KD Nuggets
Abbott Analytics
Airport Butler--OnTime Flight Stats
Predictive Analytics World - Feb09
Alltop - Top Data Mining News

DATA MINING CONFERENCES

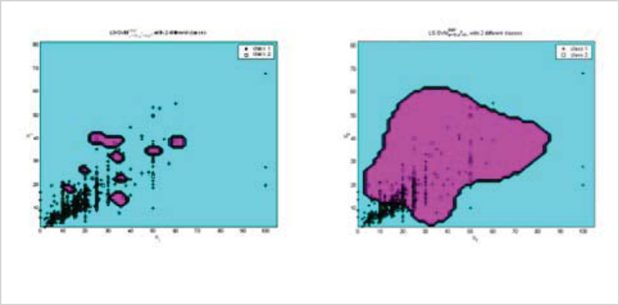
KDNuggets List of Meetings

LABELS

algorithms (2) analytics (1) art (1) baseball (3) big data (2) bioinformatics (1) books (3) business (1) Business analytics (2) business intelligence (4) business objectives (1) business understanding (1) career (1) careers (1) chart (1) classification (2) competition (1) competitions (1) computer science (1) conferences (4) contest (1) CRISP-DM (1) critical junctures (1) Data evaluation (1) data mining (18) data mining books (4) data mining competition (1) data mining conferences (4) data mining contest (1) data mining data (1) data mining degree (1) data mining education (1) data mining perceptions (1) data mining software (2) data mining survey (1) data mining training (1) data mining users (1) data mining vs. statistics (1) data preparation (4) data reduction (1) data science (1) data selection (1) Data understanding (2) data visualization (2) decision trees (1) decisions (1) distributions (1) DIY (1) dm radio (1) Do (1) Do Not (1) documentation (1) Dorian Pyle (1) due diligence (1) economics (1) eigenanalysis (1) ensembles (1) error metrics (1) ethics (1) feature selection (1) flash versus substance (1) fundamentals (1) graph (1) graphics (3) graphing (1) graphs (1) hate (1) humor (2) idempotent (1) idempotnence (1) insurance (1) integration (1) job (1) jobs (2) Kaggle (1) KDD (2) link analysis (2) literature (1) logistic regression (1) Mac-PC compatibility (1) machine

we want models to behave reasonably for data in between data that exists in the training data.

Consider the second figure below showing decision boundaries for two models built from a data set derived from the famous KDD Cup data from 1998. The two dimensions in this plot are Average Donation Amount (Y) and Recent Donation Amount (X). This data tells the story that higher values of average and recent donation amounts are related to higher likelihoods of donors responding; note that for the smallest values of both average and recent donation amount, at the very bottom left of the data, the regions are colored cyan.



Both models are built using the Support Vector Machines (SVM) algorithm, but with different values of the complexity constant, C. Obviously, the model at the left is more complex than the model on the right. The magenta regions represent responders and the cyan regions represent non-responders.

In the effort to be more accurate on training data, the model on the left creates closed-decision boundaries around any and all groupings of responders. The model at the right joins these smaller blobs together into a larger blob where the model classifies data as responders. The complexity constant for the model at the right gives up accuracy to gain simplicity.

Which model is more believable? The one on the left will exhibit strange interpolation properties; data in between the magenta blobs will be called non-responders, sometimes in very thin regions between magenta regions; this behavior isn't smooth or believable. The model at the right creates a single region of data to be classified as a responder and is clearly better than the model at the left.

learning (1) missing data (2) missing values (2) model assessment (1) model performance (1) model selection (3) modeling (3) morality (1) MSE (1) open source (1) organizational (1) orgranizations (1) outliers (1) parody (1) PCA (1) plot (1) Powerpoint (1) practice (1) practitioner (1) prediction (1) predictive (1) predictive analytics (2) predictive analytics conference (1) Predictive Analytics World (1) principal component (1) principal components analysis (1) privacy (2) problem deifition (1) programming (2) R (1) random (1) rare events (1) regular expressions (1) Rexer Analytics (1) risk (1) ROC (1) sampling (3) science (1) skills (1) software (5) spam (1) speaking (1) special values (1) statistical significance (1) statistics (5) survey (1) survey analysis (1) testing (1) text mining (2) theology (1) theorist (1) theory (2) torture (1) trendiness (1) trends (3) uplift (1) webinars (1) whimsical (1)

INSURANCE

Insured by TechInsurance

POPULAR POSTS

Similarities and Differences Between Predictive Analytics and Business Intelligence

I've been reminded recently of the overlap between business intelligence and predictive analytics. Of course any reader of this blog (or at...

ta Mining's Forgotten Step-Children

pending on whose definition one reads, the list of activities which comprise data mining will vary, but the first two items are always the...

hy Overfitting is More Dangerous than Just Poor Accuracy, Part II

Part I, I explained one problem with overfitting the data: estimates of the target variable in regions without any training data can be ...

Beware of overfitting the data and test models not just on testing or validation data, but if possible, on values not in the data to ensure its behavior, whether interpolation or extrapolation, is believable.

In part II, the problem overfitting causes for model interpretation will be addressed.

This article first appeared at the Predictive Analytics Times, http://www.predictiveanalyticsworld.com/patimes/why-overfitting-is-more-dangerous-than-just-poor-accuracy-part-i/

POSTED BY DEAN ABBOTT AT 1:20 PM

5 COMMENTS:

Anonymous said...

I would suggest that if you wish to classify a record that appears in the top left of the first figure you cannot use either of the two models shown. The model is only relevant to the data on which it has been built. Once the data you wish to classify is out of this range then the model is no longer valid.

2:14 AM

Dean Abbott said...

I agree with you that the models are only applicable to where the data was during training. Finding the gaps/empty areas in the decision space can be difficult though. It's easy to test model inputs and if all the inputs exceed their max value, you know the model has to extrapolate.

But if some of the inputs exceed and others don't, the data could still be in a good location. Or worse yet, you can have outliers interior to the range of the variables that is still not a stable place for model decisions. These are very difficult to find (remember that in multi-dimensional modeling, we can't look at the data and see these outliers). The second figure is a good example of this. Finding the right level of model complexity in these situations is very important.

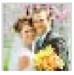
9:10 AM

acking Model Performance Over Time

ntext Most introductory data mining texts include substantial coverage of model testing. Various methods of assessing true model perform...

Why normalization matters with K-Means

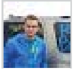
A question about K-means clustering in Clementine was posted here . I thought I knew the answer, but took the opportunity to prove it to mys...



pickettbd said...

This comment has been removed by the author.

6:34 AM



pickettbd said...

Your title drew me in. I certainly agree that overfitting is more dangerous than poor accuracy. I would also suggest that poor accuracy isn't very dangerous, making your assertion not terribly surprising (not to say it isn't a valid point, of course). If you create a model (or your learning algorithm does it for you) and the model performs poorly, you know it performs poorly up-front. When you know your model isn't very good, you either don't use or use it with caution, rendering it relatively harmless.

Overfitting, on the other hand, is dangerous (as you've identified) *because it can be difficult to detect*. You've referenced a few examples of how various models avoid overfitting and I believe anyone using these models must make themselves familiar with these techniques. Fine tuning parameters for various learning algorithms will, in many cases, be domain specific - requiring the analyst to take care.

I appreciated your Iris example. I agree that the QDA model is unnecessarily complex and the LDA model is much more appropriate. I appreciate your remarks about interpolation. We certainly do want our models to behave reasonably well for data in between existing data. I suppose that is the very purpose of creating a model in the first place. I must, however, respectfully disagree with the point about extrapolation. While it feels nice to have a model behave "reasonably" for points outside the range, we have no real way of measuring - even qualitatively - whether or not it is reasonable. If overfitting is a snake hiding in the grasses of your analysis awaiting the chance to poison your results, extrapolation must be some kind hungry carnivore. It's like trying to protect yourself from the snake by using a wolf as shield.

Finally, I think you've made a great point out of the KDD Cups example. Surely the SVM results on the left (with a bunch of small groups) is not very useful or intuitive. The one on the right is much more applicable. Overall, I think you're correct: overfitting is dangerous - much more dangerous than poor accuracy.

6:37 AM

Dean Abbott said...

thanks for your comments. You are correct that there is a bit of hyperbole going on with the title. The "dangerous" label would only be the case if the model is used, of course.

What I'm most uncomfortable with in this post is how to detect the problems. Yes, there are obvious visual cues and yes we can examine training/testing accuracy metrics (for consistency...but there is no agreed-upon standard for how different training/testing results can be before we suspect overfitting). Or better yet, if resampled data (like bootstrapped or cross-validated data) behaves consistently (accuracy), I'm much more confident as well.

I think what I'm really trying to get at and what I do in practice is stability. If different modeling algorithms predict data consistently then I'm more confident that the model will behave well. But I don't have a standard practice here...my methods change based on the algorithm (they can be unstable in different ways) and data size (bootstrapping 10M records doesn't help a lot unless the model itself is very very very complex).

7:03 AM

Post a Comment

LINKS TO THIS POST

Create a Link

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)